# Einführung in die Medieninformatik

Für EMI im 1. Semester beim Bart

# Text und HTML

## Zeichensätze

Zeichensatz definiert Zuordnung von Zahlen zu Zeichen.

1. **ASCII**: 7-bit-Zeichensatz (128 Zuordnungen), 0-31 Steuerzeichen. Neuere Zeichensätze sind meist abwärtskompatibel (Erste 128 Zuordnungen entsprechen ASCII)
2. **ISO-8859-1, ISO-8859-2 ... ISO-8859-16**: 8-bit-Zeichensätze (256 Zuordnungen), 0-127 ist ASCII, darüber Sonderzeichen wie deutsche Umlauts (-1: Latin1 westeuropäisch, -2: Latin2 osteuropäisch, -5: Kyrillisch, -6: Arabisch, -7: Griechischi [...], -15: Westeuropäisch mit Eurozeichen)
3. **Unicode**: Internationaler Zeichensatz. Verschiedene Standards: UCS-2 mit 2 hoch 16 Zeichen, UCS-4 mit doppelt so vielen Ebenen. Unicode 5 kennt 99k Zeichen.
   Unicode-Formate:
   - *UTF-32*: Immer 4 Bytes für ein Zeichen
   - *UTF-16*: 2 Bytes pro Zeichen mit Möglichkeit für mehr, wenn in ersten 2 Bytes spezille Surrogate-Kodierung benutzt wird
   - *UTF-8*: Wichtigstes Unicode-Format, weit verbreitet z. B. für XML. Besteht aus einem Byte pro Zeichen, das nur ASCII ist, wenn das oberste Bit gesetzt ist, ansonsten gehören die nächsten Bytes auch mit zum Zeichen, dann 2-4 Bytes. Aufbauschema: Höchstes Bit ist nicht gesetzt, wenn es nur ein ASCII-Zeichen ist und somit nur aus einem Byte besteht. Ist das höchste Bit gesetzt, gibt die Anzahl der dann folgenden Bits die 1 sind an, aus wie vielen Bytes das Zeichen besteht. Ist das erste Byte z. B. 11100000, folgen noch zwei weitere Bytes für das gleiche Zeichen.

## Braille-Schrift

Von Louis Braille 1820 für Blinde entwickelt, für viele natürliche Sprachen verfügbar. Gibt Zeichen als fühlbare Punkte auf dem Papier wieder (6 - 8 Punkte kodieren ein Zeichen). Spezielle Schreibweisen für Formeln, Musik o. ä.

## Schrift

---

Gebaut mit LaTeX